PR #21849 完整报告

sgl-project/sglang

[VLM]: allow Qwen3.5 models for encoder disaggregation

合并时间：2026-04-07 02:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21849

执行摘要

本 PR 修复了 Qwen3.5 多模态模型在编码器分离（EPD）部署中被错误拒绝的 bug，通过更新服务器参数验证列表和相关处理器逻辑，使 Qwen3.5 模型支持编码器分离，并添加了端到端测试验证。

功能与动机

根据 issue #21805，SGLang 运行时已支持 Qwen3.5 多模态模型，但服务器启动时编码器分离验证因允许列表过旧而失败，阻碍了有效的 EPD 部署。根本原因是 server_args.py 中的架构允许列表未包含 Qwen3.5 的模型类型。

实现拆解

- 服务器参数验证：在 python/sglang/srt/server_args.py 的 ENCODER_DISAGGREGATION_MODEL_ARCH_CHOICES 列表中添加 Qwen3_5ForConditionalGeneration 和 Qwen3_5MoeForConditionalGeneration。
- 编码器服务器逻辑：更新 python/sglang/srt/disaggregation/encode_server.py 和 python/sglang/srt/multimodal/processors/qwen_vl.py 中的模型类型检查，包含 qwen3_5 和 qwen3_5_moe 以处理视频元数据。
- 测试覆盖：在 test/registered/distributed/test_epd_disaggregation.py 中添加 TestEPDDisaggregationQwen35 测试类，验证图像和视频请求的 EPD 功能。

评论区精华

- 代码结构优化：gemini-code-assist[bot] 建议将允许列表从列表改为集合以提高效率，但未在 PR 中采纳。

"Since this collection is primarily used for membership testing, using a set would be more idiomatic and efficient."

- 正确性验证：ShangmingCai 询问是否需要修改 qwen3_5.py，Ratish1 解释 Qwen3.5 继承自 Qwen3VL，EPD 标志已处理，因此无需额外改动。
- 测试优化：ZhengWG 建议在 CI 中跳过测试以减少流水线时间，Ratish1 响应并添加了 skipIf 装饰器。

风险与影响

- 技术风险：验证逻辑变更可能影响其他模型兼容性；新增测试在 CI 中跳过，可能降低持续验证覆盖率；多模态处理器修改需确保与现有模型行为一致。
- 影响范围：用户现在可以使用 Qwen3.5 模型进行编码器分离部署，提升多模态应用灵活性；系统扩展了 EPD 支持的模型范围；团队需维护新增测试并关注类似更新。

关联脉络

与历史 PR 21921（异构 TP 测试和文档）和 22111（LTX2.3 模型支持）相关，展示了多模态模型集成和测试扩展的持续演进趋势。